

Detection and Architecture of Small Heat Shock Protein Monomers

Pierre Poulain*, Jean-Christophe Gelly, Delphine Flatters*

DSIMB, Inserm UMR-S 665 and Université Paris Diderot - Paris 7, INTS, Paris, France

Abstract

Background: Small Heat Shock Proteins (sHSPs) are chaperone-like proteins involved in the prevention of the irreversible aggregation of misfolded proteins. Although many studies have already been conducted on sHSPs, the molecular mechanisms and structural properties of these proteins remain unclear. Here, we propose a better understanding of the architecture, organization and properties of the sHSP family through structural and functional annotations. We focused on the Alpha Crystallin Domain (ACD), a β sandwich fold that is the hallmark of the sHSP family.

Methodology/Principal Findings: We developed a new approach for detecting sHSPs and delineating ACDs based on an iterative Hidden Markov Model algorithm using a multiple alignment profile generated from structural data on ACD. Using this procedure on the UniProt databank, we found 4478 sequences identified as sHSPs, showing a very good coverage with the corresponding PROSITE and Pfam profiles. ACD was then delimited and structurally annotated. We showed that taxonomic-based groups of sHSPs (animals, plants, bacteria) have unique features regarding the length of their ACD and, more specifically, the length of a large loop within ACD. We detailed highly conserved residues and patterns specific to the whole family or to some groups of sHSPs. For 96% of studied sHSPs, we identified in the C-terminal region a conserved I/V/L-X-I/V/L motif that acts as an anchor in the oligomerization process. The fragment defined from the end of ACD to the end of this motif has a mean length of 14 residues and was named the C-terminal Anchoring Module (CAM).

Conclusions/Significance: This work annotates structural components of ACD and quantifies properties of several thousand sHSPs. It gives a more accurate overview of the architecture of sHSP monomers.

Citation: Poulain P, Gelly J-C, Flatters D (2010) Detection and Architecture of Small Heat Shock Protein Monomers. PLoS ONE 5(4): e9990. doi:10.1371/journal.pone.0009990

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received: December 17, 2009; **Accepted:** March 10, 2010; **Published:** April 7, 2010

Copyright: © 2010 Poulain et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Financial support was provided by grants from the French Ministry of Research, the Paris Diderot - Paris 7 University, the National Institute for Blood Transfusion (INTS) and the Institute for Health and Medical Care (INSERM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pierre.poulain@univ-paris-diderot.fr (PP); delphine.flatters@univ-paris-diderot.fr (DF)

Introduction

Small Heat Shock Proteins (sHSPs) belong to the large superfamily of protein chaperones. More precisely, they present an ATP-independent chaperone-like activity since they bind (but not refold) non-native proteins [1–3]. Under stress conditions (high temperature, oxidative stress, etc.), they thereby prevent the irreversible aggregation of misfolded proteins. Their key roles in the cell have been demonstrated through the description of numerous mutations in human sHSPs found to be involved in severe pathologies (desmin-related myopathy [4], neurodegenerative diseases, distal hereditary motor neuropathy, cataract or tumors) [5–7]. sHSPs are ubiquitous proteins found from few in archaea, bacteria or yeast [8], to a dozen in humans [9–11] and more than 15 in plants [12]. In higher multicellular eukaryotes, each sHSP has a specific subcellular localization and/or tissue distribution [13].

An sHSP monomer has a molecular weight between 15 and 40 kDa and shares a conserved domain of 80 to 100 amino acids called the Alpha Crystallin Domain (ACD) [14,15]. Most sHSPs are structurally organized as large oligomers. Upon cell stress, they adjust their oligomeric state to bind misfolded substrates [3,5,16–8]. Despite their critical role in the cell, the mechanisms of sHSP

function are not well known and obtaining experimentally resolved structures for this family is still challenging [17,19]. This difficulty is due to the high plasticity of sHSP quaternary structures since these proteins are able to change their oligomeric state under different conditions, to exchange their subunits or to show polydisperse properties [3,20].

To date, only six sHSP structures have been resolved at atomic resolution: namely HSP 16.5 (PDB 1SHS) [16], HSP 16.9 (PDB 1GME) [17], TSP 36 (PDB 2BOL) [18], HSP A (PDB 3GLA) [21], HSP 20 (PDB 2WJ5) [22] and α B crystallin (PDB 2WJ7) [22]. These homopolymeric structures display a large variation in oligomer states (size and shape), but their monomers all share a common β sandwich fold composed of two sheets with respectively four (termed β_2 , β_3 , β_8 , β_9) and three (termed β_4 , β_5 , β_7) β strands. The second β sheet is characterized by a large L57 loop, linking strands β_5 and β_7 , and involved in dimerization in HSP 16.5 [16], HSP 16.9 [17], TSP 36 [18] and HSP A [21]. The β sandwich region is flanked by the N-terminal and C-terminal regions, described as variable in both length and amino acid composition [8,14]. These structural data have highlighted that the well-conserved β sandwich fold is in fact associated with the conserved sequence region known as ACD.

In vitro studies showed that ACD is essential for the construction of dimers and higher-order assemblies, but also for the function of sHSPs [23–25]. The N-terminal region, described as disordered, is associated with poorly conserved sequences and is generally hydrophobic [20]. It is involved in substrate binding and in higher oligomeric assembly [19,26]. The C-terminal region, flexible and unstructured, shows sequence variability with a polar tendency. This region participates in stabilizing and solubilizing the oligomeric assemblies [27]. A well-conserved motif in the C-terminal region has been shown to be involved in the inter-dimer interactions of HSP 16.5 [16] and HSP 16.9 [17]. The most divergent fragments in the sequence (the L57 loop in ACD, the N-terminal and C-terminal regions) are thought to be responsible for the large variability observed in oligomeric structures. For instance, the insertion of a peptide in the HSP 16.5 N-terminal sequence leads to either larger symmetric or polydisperse assemblies [28].

In contrast to the few resolved structures, numerous sHSP sequences are available in the UniProt databank [29] and have been annotated as belonging to the small Heat Shock Protein (or HSP20) family, primarily based on sequence information. sHSPs are associated with the PROSITE [30] profile PS01031 [15]. This profile is based on a sequence alignment of a conserved domain of about 100 residues related to ACD [14,31]. Proteins identified as sHSPs are also linked to the Pfam [32] motif PF00011, which is built from an alignment of a region of 115 amino acids roughly corresponding to ACD and the C-terminal region.

In early studies of sHSPs, sequence information was extensively used to explore evolutionary analyses [14,15,33,34], to discover new human sHSPs [10] or, more generally, to link sequence, structure and/or function in the sHSP family [3,15,35]. For instance, on the basis of an alignment of 344 unique sequences, Fu and Chang identified a highly conserved P-G doublet in non-animal sHSPs [36]. In another work based on the alignment of 26 sHSP sequences, the impact of substituting non-conserved residues in the $\beta 3$ strand on the assembly or functional properties of αB crystallin was assessed [37].

These previous works were all based on limited sets of sequences (less than 350 proteins) compared by multiple sequence alignments. However, this approach is not suitable to study several thousand sHSPs since these proteins are heterogeneous in their sequence, albeit the conservation of ACD [12,15,33,34]. For this reason, no systematic delineation and analysis of the sHSP monomer with respect to its structural topology has been yet established.

Here, we used an original approach that consists in building a Hidden Markov Model (HMM) profile based on available

structures and structurally well-annotated sequences and then in enriching it using an iterative procedure. This method leads to a robust delineation and an accurate identification of ACD in sequences of the UniProt databank. We defined three regions from a structural point of view. ACD is the region delimited from the $\beta 2$ strand to the $\beta 9$ strand. The N-terminal is the region preceding the $\beta 2$ strand and the C-terminal is the region following the $\beta 9$ strand. We identified a specific fragment in the C-terminal region that we named the C-terminal Anchoring Module (CAM). This fragment starts after ACD and includes a conserved motif. Residues that follow were defined as the C-terminal tail. Finally, we indicated specific sequence properties for sHSPs belonging to taxonomic groups such as plants and animals or previously studied groups such as class A bacteria and class B bacteria [36,38]. With this structure-based procedure, we analyzed an exhaustive set of sHSP sequences, characterized their architectural features and established sequence/structure relationships.

Results

The global procedure defined in this study is plotted in Fig. 1 and was based on three steps: (1) the constitution of an initial multiple structure/sequence alignment of known ACD structures and well-annotated sequences, (2) the construction of a Hidden Markov Model (HMM) profile of ACD based on the previous alignment and enriched with an iterative procedure using the HMMER software, and finally, (3) an exhaustive survey of the UniProt databank to detect ACDs.

Multiple structure/sequence alignment

In step (1), the 11 ACDs used as references are representative of the sHSP family (animals, plant, archae, bacteria, fungi). The nomenclature of the β strands (from $\beta 1$ to $\beta 10$) in the β sandwich fold was defined relative to the first sHSP structure to be elucidated, HSP 16.5 [16]. Here, we defined ACDs from the beginning of the $\beta 2$ strand to the end of the $\beta 9$ strand. The $\beta 2$ strand is the first structural element shared by all available X-ray structures, except for mammals HSP 20 and αB crystallin (in only one monomer of the dimeric structure). It is also involved in the dimerized structure of HSP 16.5, HSP 16.9 and HSP A. The $\beta 9$ strand is the last strand of ACD and is found in all known ACD structures. The $\beta 1$ strand, not assigned in other structures, is localized in the N-terminal region for our definition of ACD. Strands $\beta 6$ (in the L57 loop) and $\beta 10$ (corresponding to the conserved motif in the C-terminal region) were only assigned as β strands in the context of subunit interactions (i.e. for HSP 16.5 and HSP 16.9 structures). Fig. 2A shows a schematic diagram of the

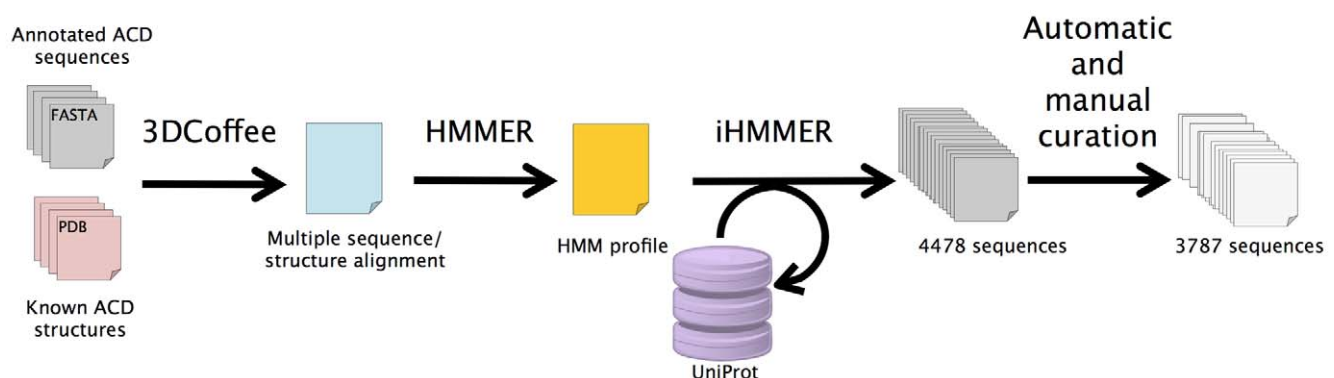


Figure 1. Flowchart of the pipeline set up in this work.

doi:10.1371/journal.pone.0009990.g001

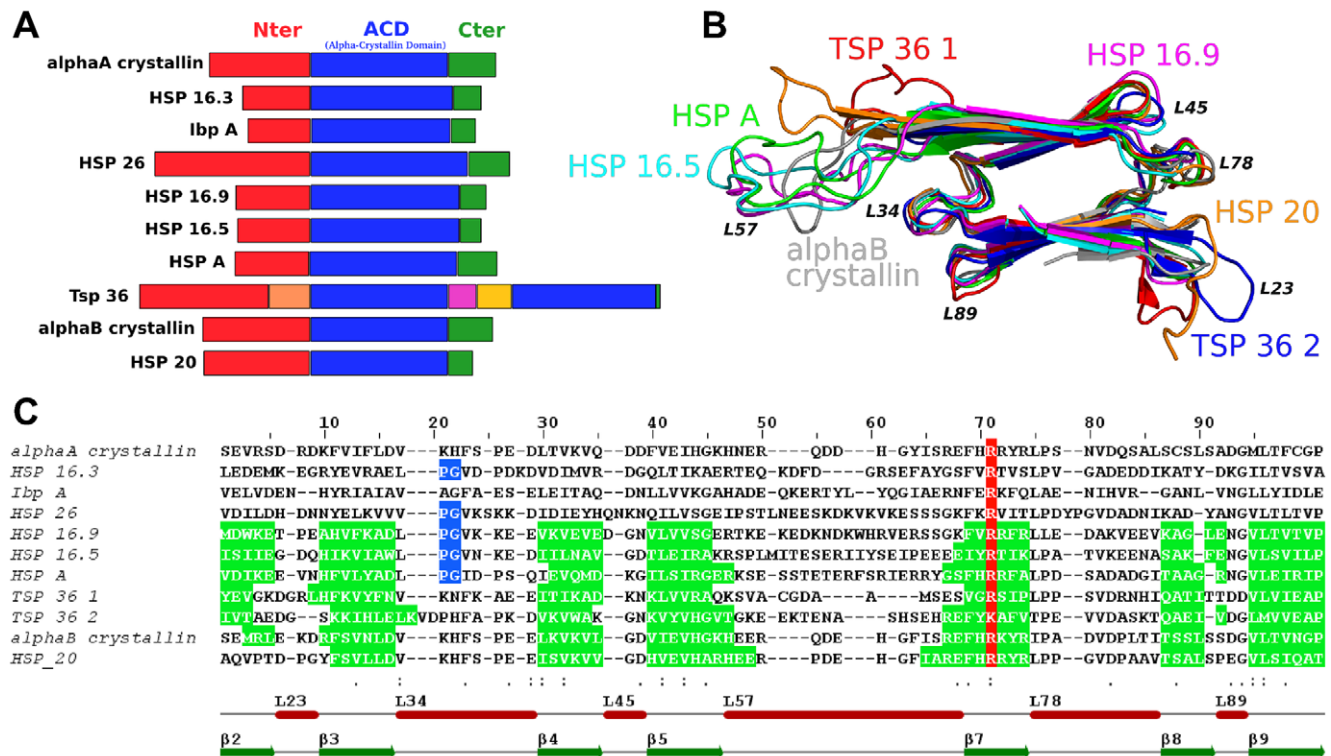


Figure 2. Reference sHSP sequences, superposition of ACD structures and initial multiple structure/sequence alignment. (A) Scheme of the 10 sequences used for the construction of the profile with delineation of the three regions: the N-terminal region in red, ACD in blue and the C-terminal region in green. For the TSP 36 sequence that contains two ACDs, linkers [18] are shown in orange, pink and yellow (B) Structural alignment of known ACDs with in-between-strand loop annotation, 3D representation made using PyMOL [59] (C) Multiple structure/sequence alignment from 3DCoffee on the ACD region. The secondary structure assignment is shown in green, β strands are annotated as green arrows and loops as red bars. ACD is subdivided into three parts: the β 2– β 5 zone, the L57 loop and the β 7– β 9 zone. The conserved arginine [15,40] is shown in red and the P-G doublet [15,36] in blue. Alignment representation performed using Jalview [60]. doi:10.1371/journal.pone.0009990.g002

reference sequences. As reported in the literature, the length of ACD is more constant than the other two regions. Superposition of the ACD structures of the common β sandwich fold also showed that ACD length is relatively well conserved (Fig. 2B). A pairwise structural alignment of ACD resulted in a low root mean square deviation (~ 1.5 Å) demonstrating a strong structural signature in this region. In contrast, the corresponding ACD sequences show only moderate identity (25–30%). Fig. 2C presents the multiple structure/sequence alignment obtained with the 3DCoffee web server [39]. For the seven ACD structures, the β strand assignments were well superimposed. Conserved motifs or residues are also represented, such as the P-G doublet for non-animal sHSP [15,36] and the arginine associated with human pathologies in the β 7 strand [15,40]. From this alignment, we subdivided ACD into three zones (the β 2– β 5 zone, the L57 loop and the β 7– β 9 zone) to analyze separately the most variable part in ACD (*i.e.* the L57 loop).

Iterative HMM profile construction

Based on this initial structural alignment, we built in step (2) a HMM profile iteratively enriched with other sequences retrieved from the UniProt databank and identified with high confidence as sHSPs (see Materials and Methods). By doing so, we were able to detect and delineate ACD more accurately, since HMM derived from the structural alignment profile performed significantly better than HMM derived only from sequence alignments [41,42]. At the end of this procedure, the generated HMM profile was named ACDP09.

ACD detection

Finally, in step (3), 4478 UniProt sequences showed a positive match using the ACDP09 profile (with E-value $\leq 10^{-3}$).

sHSP identification and region delineation

We assessed the quality of our results by comparing the retrieved sequences and the proteins annotated as sHSP in UniProt. Results are shown in Table 1. Four annotations were tested, namely family: “small heat shock protein”, family: “HSP20”, “prosite PS01031”, and “pfam PF00011”. The annotations “small heat shock protein” and “HSP20” are common names for sHSPs and are in fact synonymous because they were associated with the same sequences and 95% of these sequences were included in our dataset. The “prosite PS01031” and “pfam PF00011” annotations were assigned to more sequences and 92% and 91%, respectively, of them were also found in our dataset.

Sequences not retrieved by our method but labeled as sHSP with one of the four annotations above were too small (50 residues or less) or defined as fragment (with a potentially incomplete ACD pattern). Sequences identified by us as sHSP but not such annotated are sequences recently integrated in UniProt and not yet curated (either manually or automatically). Two months after this initial annotation comparison, about 350 more sequences were indeed labeled as sHSP in UniProt. In conclusion, the ACDP09 profile is in good agreement with known profiles integrated in PROSITE or Pfam databanks. It is more specific than the latter

Table 1. Overlap between sHSP annotated sequences in UniProt and sequences detected with ACDP09.

annotation	number of UniProt sequences	% of sequences in common with ACDP09 results
family:"small heat shock protein"	4140	95
family:"HSP 20"	4140	95
"prosite P501031"	4499	92
"pfam PF00011"	4377	91

doi:10.1371/journal.pone.0009990.t001

two, since incomplete ACD were not allowed in our methodology. The sensitivity is at least equivalent to the other profiles because clearly identified and annotated sHSPs were detected with ACDP09. Furthermore, the accurate ACD detection induced the delineation of the different regions of sHSP sequences and the different zones within ACD.

Finally, from the initial 4478 retrieved sequences, we selected complete proteins with only one ACD and a high level of protein existence evidence (see Materials and Methods section). This final dataset (named sHSPdata09 and provided as Dataset S1) included 3787 sequences of which 549 (15%) were reported in animals, 688 (18%) in plants, 612 (16%) in class A bacteria (bacA), 206 (6%) in class B bacteria (bacB), 123 (3%) in fungi, 85 (2%) in other eukaryotes (other), 1378 (36%) in bacteria that were neither class A nor class B bacteria (bacOther) and 146 (4%) in archaea (for group definition, see Materials and Methods). We analyzed in detail the full sHSPdata09 dataset and the four largest and most homogeneous groups, namely animals, plants, bacA and bacB. These four groups represent 55% of sHSPdata09. From the delineation of ACD, we easily determined the flanking N-terminal and C-terminal regions. Our dataset shows that the length of the N-terminal region varied greatly (with an average length of 53 ± 35 residues), even within each group. The N-terminal region is a highly heterogeneous region for which structural information is often missing (except for HSP 16.9 and TSP 36 structures). Thus, it still remains difficult to build a relevant multiple alignment or profile to identify conserved residues or motifs for this region.

The ACD region

The ACD region is the structural basis of the sHSP family and is well described in the literature [14,23]. We found an average ACD length of 90 ± 10 residues, a value that corroborates previous studies [5,31]. The length distribution of ACD is presented in Fig. 3A for sHSPdata09 sequences. It exhibits three peaks in a very narrow range, as 90% of sequences have an ACD length of between 82 and 100 residues. However, very large differences can occur between distributions of specific groups and are shown for animals, plants, bacA and bacB in Fig. 3A. For animals, the ACD length distribution was centered at 83 residues. Particularly, the subset of the α crystallin sHSPs (13% of animals) had an ACD length of exactly 83 residues, although the full sequence length displayed more heterogeneity. In contrast, the ACD length distribution of plants and bacA shows a peak at 90 and 86 residues respectively. For bacB, the distribution shows a broad peak at 89 residues. These results are in line with the differentiation of bacA and bacB [38]. To explore ACD further, we subdivided this region into three zones: the $\beta 2$ – $\beta 5$ zone, the L57 loop and the $\beta 7$ – $\beta 9$ zone. The length distribution of the three

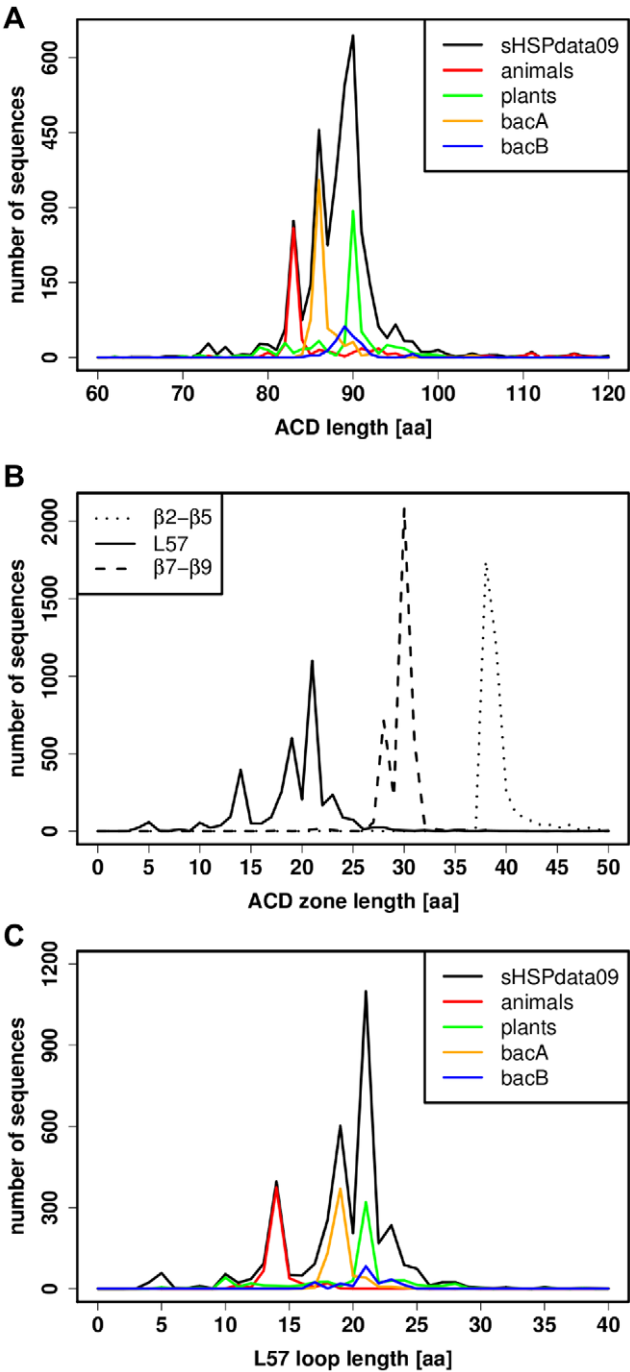


Figure 3. Length distributions of ACD and ACD zones. (A) Length distribution of ACD for sHSPdata09 sequences in black, animals in red, plants in green, bacA in orange and bacB in blue. Sequence lengths correspond to a number of amino acids (aa). (B) Length distribution of zones (inside the ACD) for sHSPdata09 with the $\beta 2$ – $\beta 5$ zone in dotted line, the L57 loop in solid line and the $\beta 7$ – $\beta 9$ zone in dashed line (C) Length distribution of the L57 loop for sHSPdata09 in black, animals in red, plants in green, bacA in orange and bacB in blue. doi:10.1371/journal.pone.0009990.g003

zones is represented in Fig. 3B. For the $\beta 2$ – $\beta 5$ zone, the distribution is centered at 38 residues. For the L57 loop, the distribution shows a first peak at 14 residues and a second, larger peak, at 21 residues. Finally, for the $\beta 7$ – $\beta 9$ zone, the distribution shows two peaks at 28 and 30 residues. The L57 loop was the

smallest zone in ACD but it was responsible for most of the variability in ACD length. A closer look at the length distribution of the L57 loop (Fig. 3C) clearly shows that animal sHSPs are generally associated with a very small L57 loop (13 residues on average). Other groups (plants, bacA and bacB) had a larger L57 loop, with a less uniform distribution for bacB and an intermediate L57 loop length for bacA. Among animals, the α crystallin sHSPs had a β 2– β 5 zone, L57 loop and β 7– β 9 zone lengths of 38, 14 and 31 residues, respectively. In addition, hydropathy scores [43] for the three zones clearly confirmed the more hydrophilic nature of the L57 loop compared to the β 2– β 5 and β 7– β 9 zones [44] (data not shown).

The logo representation of ACD in Fig. 4 highlights the most conserved residues that match the positions of the ACDP09 profile. For sHSPdata09 (Fig. 4A), some residues or motifs were highly conserved in the family, for instance, G/A at the end of the β 5 strand, F-X-R in the β 7 strand, L-P/A at the beginning of the L78 loop, N/D-G-hydrophobic-L between the L89 loop and the β 9 strand. These residues were mainly in the β 7– β 9 zone. We also noticed the presence of conserved charged residues at the end of the L34 loop and in the first half of the L57 loop.

Other positions also appeared to be highly conserved such as G19, L34 and A78, but were not common properties to all sHSPs. Moreover, many were group-specific as shown in the logo representation of animals, plants, bacA and bacB (Fig. 4B to E). For example, G19, L34 and A78 residues were not predominant in animals. Instead, specific residues or motifs such as F11, D-V-X-X-F-X-P, K28 in the β 3– β 4 zone or G-K-H-E-E-R/K in the very beginning of the L57 loop, Y65 (rather than F65) in β 7 or a serine-rich motif in β 8 (S-S/T-L-S) were observed in this group.

The G19 residue actually belongs to the P-G doublet involved in the dimer interface of two subunits (for HSP 16.5, HSP 16.9 and HSP A structures). In a study on 344 sHSPs, Fu *et al.* [36] showed that the P-G doublet is highly conserved in most of the non-animal sHSPs. The L34 loop (annotated in Fig. 4 in positions 17 to 24) generally starts with an aliphatic residue and ends with two (usually negatively) charged residues. As in ref. [36], the P-G doublet is surrounded by hydrophobic residues (Fig. 4). In animals, this motif was missing and was replaced with charged amino acids [36]. The hydrophobic residue in position 20 was usually a F (as for bacA) and a P residue was found in position 22 (Fig. 4B). The plant and bacB groups clearly showed the P-G doublet (Fig. 4 C and E), whereas the bacA group showed a highly conserved A-G doublet, with neighboring residues being mainly V and F (Fig. 4D). Although we used a different methodology and had ten times as many sequences, our results are perfectly in line with the findings of Fu *et al.* [36]. Furthermore, the bacA group showed a unique, intermediate feature. Based on their highly conserved A-G motif, bacA sHSPs were similar to plants and bacB. However, the P-to-A substitution in the original P-G doublet was unique and exclusive (97% of bacA sequences possessed the A-G doublet compared to 1% for P-G). Finally, hydrophobic neighboring residues of the doublet were mainly V and F, as for animals. In addition, sequences in animals and bacA shared several other common residues such as H53, G55, R59 in the second half of the L57 loop.

In bacA, the sequence logo representation showed several well-conserved motifs in β 2 (Y-N-I/V-E), β 3 and the L34 loop (Y-R-I-X-X-A-V/L-A-G-F), and in the beginning of L78 loop (L-A-D-E). In contrast to other groups, the sequence logo revealed a very well-conserved second half of L57 loop, a shortest L78 loop (no residue detected in positions 75 and 76) and, a specific β 9 end (D/E-L-X). BacA sequences were also characterized by the absence of proline residues that were usually conserved in other groups (positions 18, 68, 91) and were replaced by an alanine (at positions 18 and 68).

Conserved alanine residues are specific to this group, such as A16 (replaced by a negatively charged D/E residue in other groups) or A57. Sequences in plants often shared several characteristics with bacB (such as residues D2 and E5 in β 2, motifs G-E-R and R-X-E-R-X-X-G in the L57 loop) or, more generally, with non-animal sequences (such as P/A-G doublet in the L34 loop, L34 in β 5 or A78 in β 8).

To summarize, the ACD region is the hallmark of the sHSP family, both from a sequence and a structural points of view. We used this feature to fetch sHSP sequences through the UniProt databank. Nevertheless, we revealed that ACD also has unique features that are specific to some groups.

The C-terminal Anchoring Module (CAM)

A conserved motif located in the C-terminal region [15] is directly involved in the oligomerization process [45] and is interacted with the hydrophobic groove of ACD in HSP 16.5 and HSP 16.9 structures [16,17]. This motif is also referenced in the literature as I-X-I [15], I/V-X-I [46], I-X-I/V [17] or I/V-X-I/V [18] patterns. The I/V-X-I/V motif was found in 90% of sHSPdata09. As I and V are aliphatic residues, we also searched for an extended I/V/L-X-I/V/L motif that we found in 96% of our dataset. This motif, found in a polar region, appeared to be an important characteristic of the sHSP family. The logo representation of the I/V/L-X-I/V/L motif in Fig. 5A shows the predominance of I residues at both extremities of the motif. However, at the central position, we could not see a clear signal of most frequent residues. We therefore give the logo representation of the motif for all groups separately (Fig. 5B to E). The central position for animals was usually a proline, which was not found in other groups (except for bacA, after A and E residues), and sometimes a glutamate (often observed in other groups). The bacA group showed predominant A/E residues in the central position and I in the first and last positions.

The strong presence of the I/V/L-X-I/V/L motif in the C-terminal region and its anchoring role in sHSP oligomers led us to study the length of the fragment from the end of ACD to the end of the motif that we defined as the C-terminal Anchoring Module (CAM). The last part of the C-terminal region, located after CAM was called the C-terminal tail (Cter tail). For sHSPdata09, CAM had an average length of 14 residues and 85% of the sequences with the motif had a CAM length of 14 ± 3 residues (Fig. 5F). All studied groups showed a similar CAM length distribution. However, drosophila and tick sHSPs presented a specific CAM length of 25–26 residues that appeared to be a strong characteristics of these two species. The Cter tail appeared rather short with 65% of sequences having a Cter tail smaller than 5 residues (Fig. 5G). Plants and bacB shared this feature, whereas animals showed an average Cter tail length of 10 residues and bacA had either short or long Cter tails.

Discussion

Detection of sHSPs with the ACDP09 profile

In the present work, we built an HMM profile dedicated to the identification and the structural annotation of protein sequences belonging to the sHSP family. The originality of our procedure lies in combining an initial profile solely based on all available structural data with an iterative procedure. In theory, this approach could be easily transposed to other protein families with low sequence identity, few available structures but with a strong structural pattern.

Our profile was initially built on a multiple sequence/structure alignment of 11 structurally well-characterized ACDs. From a

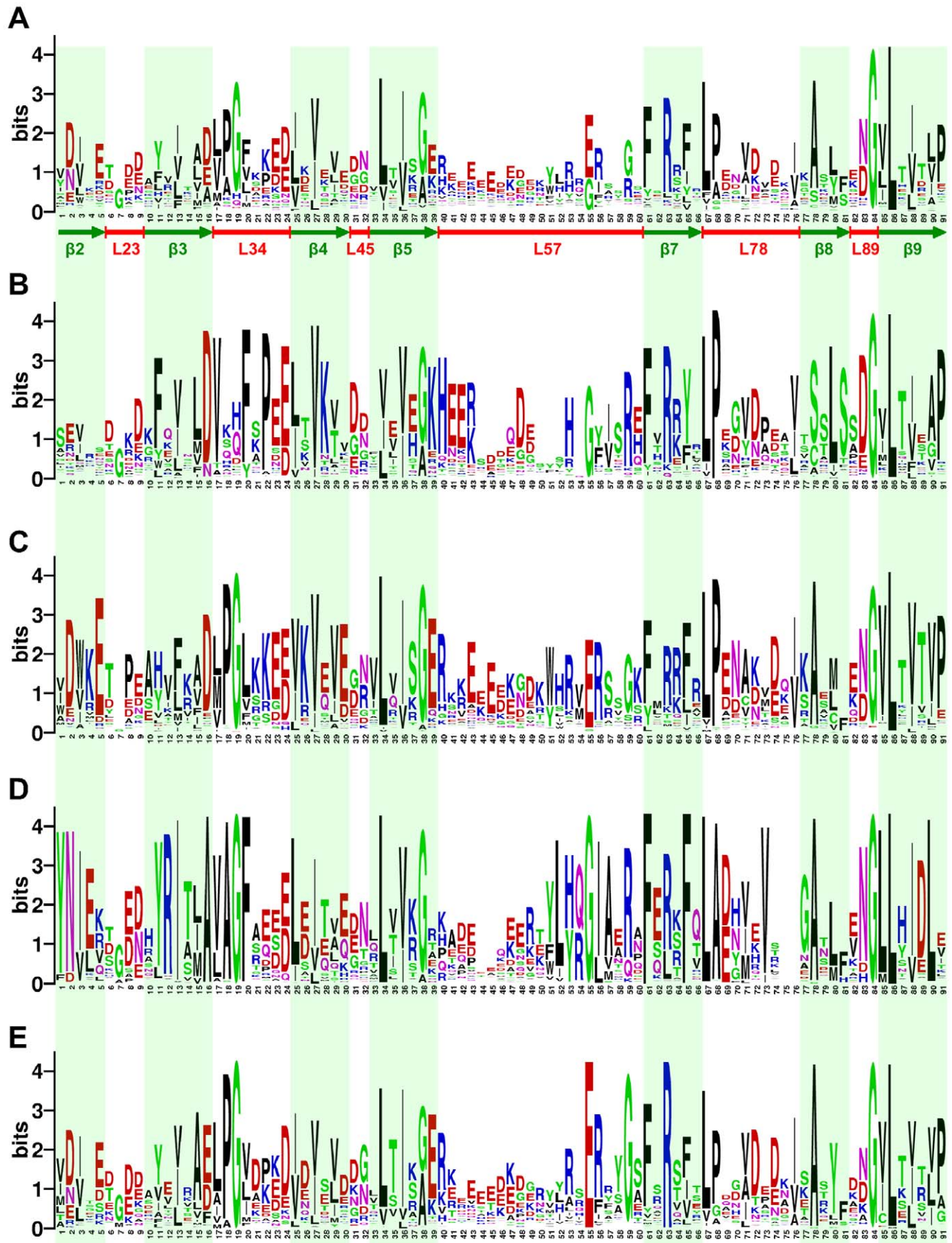


Figure 4. Logo representation of the ACD profile. The height of each letter is proportional to the information content of residues that matched ACDP09 profile positions. Amino acids are colored according to their chemical properties: acidic (D,E) in red, basic (K,R,H) in blue, polar (G,S,T,Y,C) in green and (N,Q) in purple, hydrophobic (A,V,L,I,P,W,F,M) in black. Logo representation for (A) sHSPdata09, (B) animals, (C) plants, (D) bacA and (E) bacB. The β sheets and loops are annotated in green and red, respectively.
doi:10.1371/journal.pone.0009990.g004

structural standpoint, we defined ACD as the region delimited from the β 2 strand to the β 9 strand. The ACDP09 profile fetched 4478 sequences from UniProt that were consistent with current sHSP annotations. These sequences were all characterized by the presence of one complete ACD and, after curation, 3787 proteins were analyzed as the sHSPdata09 dataset. To our knowledge, this was the first time that an exhaustive study of more than one thousand sHSPs was conducted.

Architecture of the monomer and common specificities to the sHSP family

Based on ACD detection, we deduced the architecture of sHSP monomers. We clearly delineated the three structural regions (N-terminal, ACD, C-terminal) in each sequence.

As reported in the literature, the N-terminal region is the most heterogeneous region in terms of length and sequence, making it difficult to extract any general properties of sHSPs [8,15,20].

The analysis of ACD, known as the signature of the family, revealed strong specificities shared by the sHSPdata09 sequences

identified in this study. We found that the length of ACD is 90 ± 10 residues and is highly dependent on the L57 loop length whereas the average length of the β 2– β 5 and β 7– β 9 zones is conserved. The logo representation of ACD, where positions refer to the ACDP09 profile, displays residues or motifs that are very conserved in sHSPs. Interestingly, these residues are mainly localized along the β 7– β 9 zone (β 7 strand, L78 loop, L89 loop/ β 9 strand). The motif identified in the β 7 strand contains the arginine residue associated with human pathologies [4], which is clearly conserved even on a set of 3787 sHSP sequences. Other residues observed at this position in much lower frequencies are L, K and Q. In the β 2– β 5 zone, we found alternating hydrophobic residues specific to β strands (β 4 and β 5) and a terminating G/A residue. At last, the beginning of the L57 loop is marked by charged residues followed by some isolated conserved residues.

In the C-terminal region, 96% of the sHSPdata09 sequences have the motif I/V/L-X-I/V/L, confirming that this motif is a common characteristic of sHSPs. Interestingly, the distance from the end of ACD to the end of the motif is well conserved. We

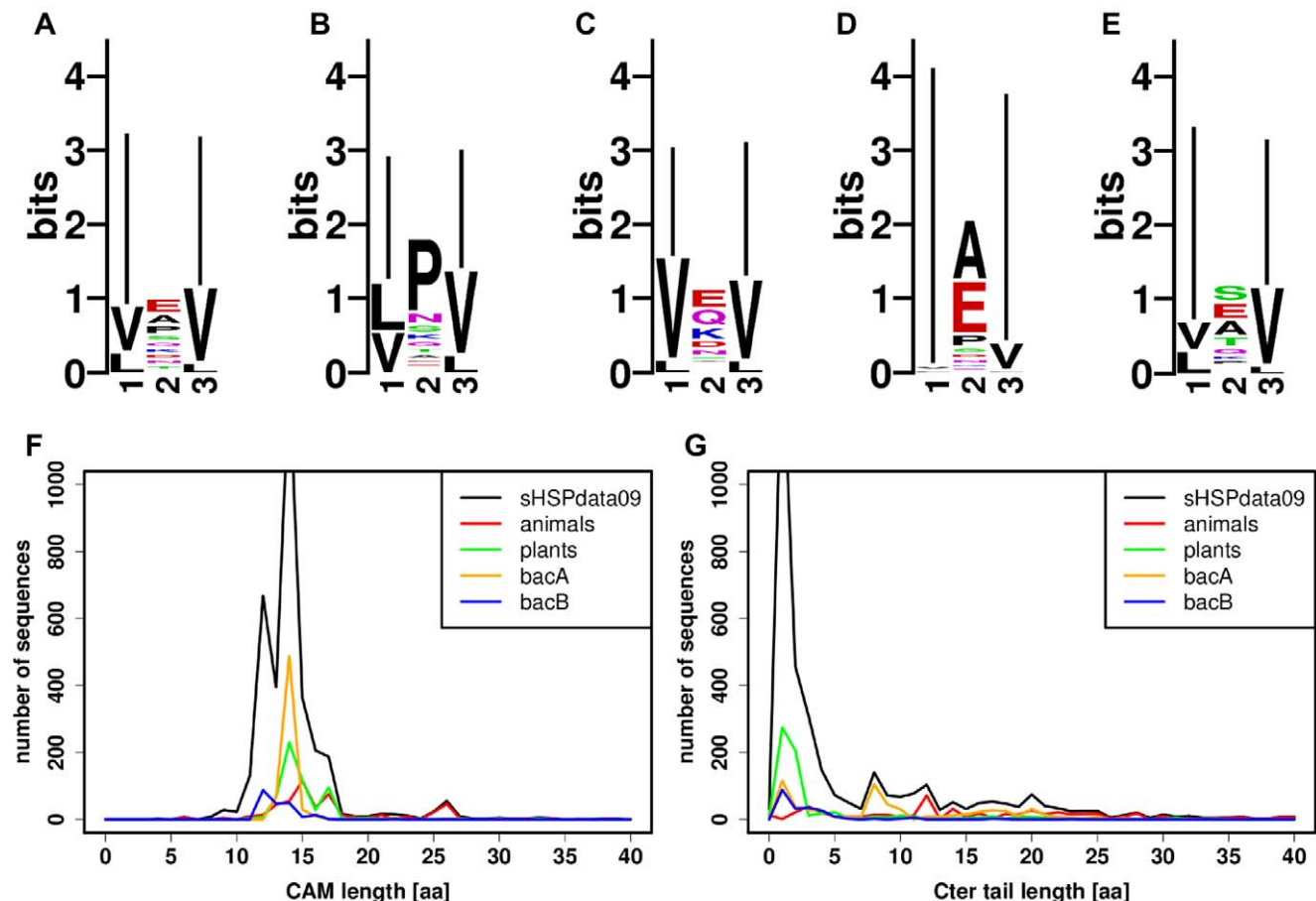


Figure 5. Logo representation of the I/V/L-X-I/V/L motif and length distributions of CAM and Cter tail. Logos are displayed for (A) sHSPdata09, (B) animals, (C) plants, (D) bacA and (E) bacB. The length distributions of CAM (F) and Cter tail (G) are computed for sHSPdata09 (in black), animals (in red), plants (in green), bacA (in orange) and bacB (in blue). The tops of the sHSPdata09 length distributions are truncated for reasons of clarity. Sequence lengths correspond to a number of amino acids (aa).
doi:10.1371/journal.pone.0009990.g005

named this fragment the C-terminal Anchoring Module (CAM). This result demonstrates that CAM has a more constrained length than what is commonly supposed on the localization of the motif in the C-terminal region [14–16]. Consequently, the variability of this region is mainly due to the Cter tail (residues following CAM). The existence of CAM supports the critical role of the I/V/L-X-I/V/L motif in the stabilization of sHSP assemblies [17,23], as illustrated by HSP 16.5 and HSP 16.9 oligomeric structures.

Divergence between sHSP groups

In this study, we focused on homogeneous groups having a sufficient number of sequences (at least 200). Four groups were considered: animals, plants, bacA and bacB. Detailed analyses highlighted group specificities that could constitute a first step toward a classification procedure (as illustrated in ref. [35]). Three distinct peaks in the ACD length distribution are associated with sequences in animals (83 residues), bacA (86 residues) and plants (90 residues). These characteristic ACD lengths are directly related to the L57 loop lengths, where animals are mainly associated with the shortest L57 loop (13 residues). The logo representations built for each group illustrate specificities in terms of structural zones or residues. Sequences in animals are distinguished from non-animals in structural elements involved in function or in the oligomerization process. Animal sequences specifically show a poorly conserved β 2 strand, and conserved residues/motifs in the β 3 strand, the L34 loop, the β 4 strand, the very beginning of the L57 loop and in the β 7 and β 8 strands. Interestingly, the β 2 strand is not always seen in mammalian structures [22], but is involved in the dimerization of non-mammalian structures (*i.e.* HSP 16.5, HSP 16.9 and HSP A). The fragment from the β 3 strand to the β 4 strand is functionally relevant because associated with substrate binding in mammalian α A and α B crystallins [24,37,47]. Finally, the β 7 and β 8 strands are involved in the association of monomers [22] or dimers [16,17], respectively. Moreover, animals share several residues with bacA (in the L34 loop and in the second half of the L57 loop), whereas plants share several characteristics with bacB (in the β 2 strand and in the L57 loop). BacA sequences display their own characteristics with specific conserved motifs (in the β 2, β 3 and β 9 strands and in the second half of the L57 loop), the shortest L78 loop of all groups and a different usage of proline and alanine amino acids. The conserved proline residues in other groups are often replaced by an alanine in bacA (the A-G and L-A doublets in the L34 and L78 loops, respectively), and conserved alanine residues in bacA are not seen in other groups (*e.g.* last residue in the β 3 strand).

An analysis of the C-terminal region also revealed group specificities. A strong group preference for the X residue in the I/V/L-X-I/V/L motif was observed: animals and bacA preferentially show I/L/V-P-I/V/L and I-A/E-I motifs respectively. To go further, a P residue at the central position is a signature of animals whereas an E residue is related to non-animals. The Cter tail (zone following the I/V/L-X-I/V/L motif) is generally very short in plants and bacB (less than three residues on average), but can be significantly longer in animals (10 residues on average) and bacA (8 residues). Indeed, this Cter tail has been described as a very flexible zone in NMR studies of α A and α B crystallins [48].

In conclusion, this work provides new insights on the structural organization of sHSP monomers. Structural elements identified as essential to the oligomerization process are associated with specific sequence properties in each group. Particularly, animals are characterized by a short L57 loop, a poorly conserved β 2 strand and the absence of the P-G doublet in the L34 loop. All these structural elements are known, however, to be involved in dimer formation in the non-animal HSP 16.9, HSP 16.5 and HSP A

structures. Differences between animal and non-animal sHSPs are confirmed by the recently published crystallographic structures of two mammalian sHSPs (and also by the worm TSP 36 structure) that present a dimer organization distinct from the previous known dimers found in plant, archaea or bacteria [22]. Finally, we show the relevance of the iterative HMM approach based on structural data, learned on sequences but aimed toward elucidating structural/functional properties of sHSPs.

Materials and Methods

Selection of the initial dataset and structural delineation of ACD

Seven ACDs are available at atomic resolution from crystallographic structures of sHSPs: HSP 16.5 (PDB 1SHS [16], UniProt Q5773) from the archae *Methanococcus jannaschii*, HSP 16.9 (PDB 1GME [17], UniProt Q41560) from the plant *Triticum aestivum*, HSP A (PDB 3GLA [21], UniProt Q8PNC2) from the bacterium *Xanthomonas axonopodis* pv. *Citri*, TSP 36 from the worm *Taenia saginata* (PDB 2BOL [18], UniProt Q7YZT0) with 2 ACD, HSP 20/HSP B6 (PDB 2WJ5 [22], UniProt P97541) from the mammal *Rattus norvegicus* and α B crystallin/HSP B5 (PDB 2WJ7 [22], UniProt P02511) from *Homo sapiens*. The structures deposited in the Protein Data Bank are either the entire complex (24-mer for HSP 16.5, 12-mer for HSP 16.9, 4-mer for TSP 36) or just dimers (for HSP A, HSP 20 and α B crystallin). The secondary structures from β 2 to β 9 are assigned from dssp [49], and confirmed by the literature and by pairwise structural comparisons. These structures were supplemented with four sHSP sequences structurally well-annotated with ACD delineation clearly indicated: human α A crystallin (UniProt P02489) (annotated from site-directed spin labeling [50] and conformational studies on a wildtype and mutants [51]), HSP 16.3 from the bacterium *Mycobacterium tuberculosis* (UniProt P0A5B7) (annotated from mass spectrometry and electron microscopy [52], multiple sequence alignment and structural characterization [36]), HSP 26 from the yeast *Saccharomyces cerevisiae* (UniProt P15992) (studied with cryoelectron microscopy [53]), and IbpA (UniProt P0C054) from the bacteria *Escherichia coli* (often included in multiple sequence alignment analysis [3,15,36,37]).

Sequence/structure multiple alignment

We designed a robust alignment procedure based on a mixed structural and sequence multiple alignment of manually delineated ACD regions. The multiple alignment was created using the 3DCoffee [39] web server using advanced options (slow_pair, sap_pair, fugue_pair). Our ACD reference set had seven ACDs extracted from 3D structures and four extracted from sequences. We carefully inspected the alignment to verify its consistency with the literature. More specifically, we manually checked that β sheet assignment and known motifs are aligned (*i.e.* the P-G doublet in non-animal sequences [36], the well-conserved arginine residue in the β 7 strand [25,40]). This high-quality alignment was subsequently used to create the initial Hidden Markov Model (HMM) profile with HMMER (version 2.3.2) [54,55] using standard parameters.

Iterative Hidden Markov Model profile generation

Protein sequences identified as sHSPs were found in the UniProt [29] databank (release 15.9 of October 13, 2009) that contains more than 10 million protein sequences. We made no distinction between reviewed (manually annotated, termed SwissProt) and unreviewed (automatically annotated, termed TrEMBL) proteins.

An iterative procedure (iHMMER) was established to perform a specific and sensitive search of the UniProt databank. The full process can be described by the following steps:

1. Generate the initial profile from the structural alignment.
2. Scan the UniProt databank for sHSPs with the profile and a given E-value threshold (start at 10^{-15}).
3. Filter (no ACD length greater than 100 residues) and extract detected ACD.
4. Align selected ACD sequences in the profile.
5. Generate a new profile from the alignment.
6. Calibrate the profile.
7. Increase by one order of magnitude the E-value threshold and return to step 2 until a threshold E-value of 10^{-5} has been reached.

The iHMMER procedure converged quickly to roughly 4000 sequences. After 10 iterations, a final step selected sequences with only one ACD and an ACD detection E-values of 10^{-3} or less.

Fetched sequences were then compared to sHSP sequences labeled within UniProt with at least one of the following annotations: family:“small heat shock protein, family:“HSP20”, “prosite PS01031” and “pfam PF0011”. The first two are common sHSP names, whereas the last two are profiles deposited in the PROSITE and the Pfam databanks, respectively.

Manual curation, sequence refinement, region analyses and sequence slicing

Sequences with a single ACD were selected based on their UniProt annotations (no fragment and protein existence level less or equal to 3) and then divided into three regions (N-terminal, ACD and C-terminal) from their ACD delineation. Following the classification performed in the work of Fu and Chang [36], sHSPs were grouped based on their taxonomic origin into archaea, fungi, plants, animals and other eukaryotes (termed other). Based on pairwise global alignments, bacterial sHSPs were classified as class

A bacteria (bacA), class B bacteria (bacB) [38] or neither class A or B (bacOther). From the positioning of the L57 loop in the profile, identified ACDs were then subdivided into the $\beta 2$ – $\beta 5$ zone (composed of the $\beta 2$, $\beta 3$, $\beta 4$ and $\beta 5$ strands), the L57 loop and the $\beta 7$ – $\beta 9$ zone (composed of the $\beta 7$, $\beta 8$ and $\beta 9$ strands). These three zones were analyzed in terms of length and hydropathy score (defined as the sum of the Kyte and Doolittle hydropathy indexes [43] of a given zone divided by its length). We also looked for the I/V-X-I/V [18] and I/V/L-X-I/V/L motifs in the C-terminal region with the help of regular expressions.

A sequence logo representation of the predominant amino acids was built from residues that matched profile positions. We use a modified version of WebLogo [56,57] where gaps are explicitly taken into account as a 21st amino acid but not drawn in the logo.

All statistical analysis and graphics were done using R [58].

Supporting Information

Dataset S1 List of the 3787 sequences constituting the sHSPdata09 dataset. First column is the UniProt accession, second column is the corresponding group and third column is the length of the detected ACD.

Found at: doi:10.1371/journal.pone.0009990.s001 (0.10 MB PDF)

Acknowledgments

We thank Dr. Alexandre G. de Brevern and Pr. Catherine Etchebest for helpful discussions, Christel Goudot and Benoist Laurent for preliminary works on the X residue in the I/V/L-X-I/V/L motif, Dr. Bénédicte Chommeloux for her support during the manuscript revisions and Oscar Foal for informal debates on the subject.

Author Contributions

Conceived and designed the experiments: PP DF. Performed the experiments: PP JCG. Analyzed the data: PP JCG DF. Contributed reagents/materials/analysis tools: PP. Wrote the paper: PP JCG DF.

References

1. Horwitz J (1992) Alpha-crystallin can function as a molecular chaperone. *Proc Natl Acad Sci USA* 89: 10449–10453.
2. Jakob U, Gaestel M, Engel K, Buchner J (1993) Small heat shock proteins are molecular chaperones. *J Biol Chem* 268: 1517–1520.
3. Narberhaus F (2002) Alpha-crystallin-type heat shock proteins: socializing minichaperones in the context of a multichaperone network. *Microbiol Mol Biol Rev* 66: 64–93.
4. Vicart P, Caron A, Guicheney P, Li Z, Prévost MC, et al. (1998) A missense mutation in the alphaB-crystallin chaperone gene causes a desmin-related myopathy. *Nat Genet* 20: 92–95.
5. Clark JI, Muchowski PJ (2000) Small heat-shock proteins and their potential role in human disease. *Curr Opin Struct Biol* 10: 52–59.
6. Sun Y, MacRae TH (2005) The small heat shock proteins and their role in human disease. *FEBS J* 272: 2613–2627.
7. Arrigo AP, Simon S, Gibert B, Kretz-Remy C, Nivon M, et al. (2007) Hsp27 (HspB1) and alphaB-crystallin (HspB5) as therapeutic targets. *FEBS Lett* 581: 3665–3674.
8. Haslbeck M, Franzmann T, Weinfurter D, Buchner J (2005) Some like it hot: the structure and function of small heat-shock proteins. *Nat Struct Mol Biol* 12: 842–846.
9. Fontaine J, Rest JS, Welsh MJ, Benndorf R (2003) The sperm outer dense fiber protein is the 10th member of the superfamily of mammalian small stress proteins. *Cell Stress Chaperones* 8: 62–69.
10. Kappé G, Franck E, Verschuure P, Boelens WC, Leunissen JA, et al. (2003) The human genome encodes 10 alpha-crystallin-related small heat shock proteins: HspB1–10. *Cell Stress Chaperones* 8: 53–61.
11. Kampinga HH, Hageman J, Vos MJ, Kubota H, Tanguay RM, et al. (2009) Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones* 14: 105–111.
12. Siddique M, Gernhard S, von Koskull-Döring P, Vierling E, Scharf KD (2008) The plant sHSP superfamily: five new members in *Arabidopsis thaliana* with unexpected properties. *Cell Stress Chaperones* 13: 183–197.
13. Vos MJ, Hageman J, Carra S, Kampinga HH (2008) Structural and functional diversities between members of the human HSPB, HSPH, HSPA, and DNAJ chaperone families. *Biochemistry* 47: 7001–7011.
14. de Jong WW, Leunissen JA, Voorter CE (1993) Evolution of the alpha-crystallin/small heat-shock protein family. *Mol Biol Evol* 10: 103–126.
15. de Jong WW, Caspers GJ, Leunissen JA (1998) Genealogy of the alpha-crystallin–small heat-shock protein superfamily. *Int J Biol Macromol* 22: 151–162.
16. Kim KK, Kim R, Kim SH (1998) Crystal structure of a small heat-shock protein. *Nature* 394: 595–599.
17. van Montfort RL, Basha E, Friedrich KL, Slingsby C, Vierling E (2001) Crystal structure and assembly of a eukaryotic small heat shock protein. *Nat Struct Biol* 8: 1025–1030.
18. Stamler R, Kappé G, Boelens W, Slingsby C (2005) Wrapping the alpha-crystallin domain fold in a chaperone assembly. *J Mol Biol* 353: 68–79.
19. Jaya N, Garcia V, Vierling E (2009) Substrate binding site flexibility of the small heat shock protein molecular chaperones. *Proc Natl Acad Sci USA* 106: 15604–15609.
20. Mchaourab HS, Godar JA, Stewart PL (2009) Structure and mechanism of protein stability sensors: chaperone activity of small heat shock proteins. *Biochemistry* 48: 3828–3837.
21. Hilario E, Teixeira EC, Pedrosa GA, Bertolini MC, Medrano FJ (2006) Crystallization and preliminary X-ray diffraction analysis of XAC1151, a small heat-shock protein from *Xanthomonas axonopodis* pv. *citri* belonging to the alpha-crystallin family. *Acta Crystallogr Sect F Struct Biol Cryst Commun* 62: 446–448.
22. Bagnéris C, Bateman OA, Naylor CE, Cronin N, Boelens WC, et al. (2009) Crystal structures of alpha-crystallin domain dimers of alphaB-crystallin and Hsp20. *J Mol Biol* 392: 1242–1252.
23. Sun Y, MacRae TH (2005) Small heat shock proteins: molecular structure and chaperone function. *Cell Mol Life Sci* 62: 2460–2476.
24. Bhattacharyya J, Udupa EP, Wang J, Sharma K (2006) Mini-alphaB-crystallin: a functional element of alphaB-crystallin with chaperone-like activity. *Biochemistry* 45: 3069–3076.

25. Simon S, Michiel M, Skouri-Panet F, Lechaire JP, Vicart P, et al. (2007) Residue R120 is essential for the quaternary structure and functional integrity of human alphaB-crystallin. *Biochemistry* 46: 9605–9614.
26. Aquilina JA, Watt SJ (2007) The N-terminal domain of alphaB-crystallin is protected from proteolysis by bound substrate. *Biochem Biophys Res Commun* 353: 1115–1120.
27. Lindner RA, Carver JA, Ehrmsperger M, Buchner J, Esposito G, et al. (2000) Mouse Hsp25, a small shock protein, the role of its C-terminal extension in oligomerization and chaperone action. *Eur J Biochem* 267: 1923–1932.
28. Shi J, Koteiche HA, Mchaourab HS, Stewart PL (2006) Cryoelectron microscopy and EPR analysis of engineered symmetric and polydisperse Hsp16.5 assemblies reveals determinants of polydispersity and substrate binding. *J Biol Chem* 281: 40420–40428.
29. The UniProt Consortium (2009) The Universal Protein Resource (UniProt). *Nucleic Acids Res* 37 (Database issue): D169–D174.
30. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, et al. (2008) The 20 years of PROSITE. *Nucleic Acids Res* 36 (Database issue): D245–D249.
31. Groenen PJ, Merck KB, de Jong WW, Bloemendal H (1994) Structure and modifications of the junior chaperone alpha-crystallin, from lens transparency to molecular pathology. *Eur J Biochem* 225: 1–19.
32. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue): D281–D288.
33. Plesofsky-Vig N, Vig J, Brambl RJ (1992) Phylogeny of the alpha-crystallin-related heat-shock proteins. *Mol Evol* 35: 537–545.
34. Waters ER, Vierling E (1999) Chloroplast small heat shock proteins: evidence for atypical evolution of an organelle-localized protein. *Proc Natl Acad Sci USA* 96: 14394–14399.
35. Goldstein P, Zucko J, Vujaklija D, Krisko A, Hranueli D, et al. (2009) Clustering of protein domains for functional and evolutionary studies. *BMC Bioinformatics* 10: 335.
36. Fu X, Chang Z (2006) Identification of a highly conserved pro-gly doublet in non-animal small heat shock proteins and characterization of its structural and functional roles in *Mycobacterium tuberculosis* Hsp16.3. *Biochemistry (Mosc)* 71: S83–S90.
37. Ghosh JG, Estrada MR, Houck SA, Clark JI (2006) The function of the beta3 interactive domain in the small heat shock protein and molecular chaperone, human alphaB crystallin. *Cell Stress Chaperones* 11: 187–197.
38. Münchbach M, Nocker A, Narberhaus F (1999) Multiple Small Heat Shock Proteins in *Rhizobia*. *J Bacteriol* 181: 83–90.
39. Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res* 32 (Web Server issue): W37–40.
40. Kumar LV, Ramakrishna T, Rao CM (1999) Structural and functional consequences of the mutation of a conserved arginine residue in alphaA and alphaB crystallins. *J Biol Chem* 274: 24137–24141.
41. Scheff ED, Bourne PE (2006) Application of protein structure alignments to iterated hidden markov model protocols for structure prediction. *BMC Bioinformatics* 7: 410.
42. Bernardes JS, Dávila AMR, Costa VS, Zaverucha G (2007) Improving model construction of profile HMMs for remote homology detection through structural alignment. *BMC Bioinformatics* 8: 435.
43. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157: 105–132.
44. de Jong WW, Leunissen JA, Leenen PJ, Zweers A, Versteeg M (1988) Dogfish alpha-crystallin sequences. Comparison with small heat shock proteins and *Schistosoma* egg antigen. *J Biol Chem* 263: 5141–5149.
45. Saji H, Iizuka R, Yoshida T, Abe T, Kidokoro SI, et al. (2008) Role of the IXI/V motif in oligomer assembly and function of StHsp14.0, a small heat shock protein from the acidothermophilic archaeon, *Sulfolobus tokodaii* strain 7. *Proteins* 71: 771–782.
46. Pasta SY, Raman B, Ramakrishna T, Rao CM (2004) The IXI/V motif in the C-terminal extension of alpha-crystallins: alternative interactions and oligomeric assemblies. *Mol Vis* 10: 655–662.
47. Sharma K, Kumar R, Kumar G, Quinn P (2000) Synthesis and characterization of a peptide identified as a functional element in alphaA-crystallin. *J Biol Chem* 275: 3767–3771.
48. Carver JA, Aquilina JA, Truscott RJ, Ralston GB (1992) Identification by ¹H NMR spectroscopy of flexible C-terminal extensions in bovine lens alpha-crystallin. *FEBS Lett* 311: 143–149.
49. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22: 2577–2637.
50. Koteiche HA, Mchaourab HS (1999) Folding pattern of the alpha-crystallin domain in alphaA-crystallin determined by site-directed spin labeling. *J Mol Biol* 294: 561–577.
51. Bera S, Thampi P, Cho WJ, Abraham EC (2002) A positive charge preservation at position 116 of alphaA-crystallin is critical for its structural and functional integrity. *Biochemistry* 41: 12421–12426.
52. Kennaway CK, Benesch JL, Gohlke U, Wang L, Robinson CV, et al. (2005) Dodecameric structure of the small heat shock protein Acr1 from *Mycobacterium tuberculosis*. *J Biol Chem* 280: 33419–33425.
53. White HE, Orlova EV, Chen S, Wang L, Ignatiou A, et al. (2006) Multiple distinct assemblies reveal conformational flexibility in the small heat shock protein Hsp26. *Structure* 14: 1197–1204.
54. Eddy SR (1998) Profile hidden markov models. *Bioinformatics* 14: 755–763.
55. HMMER: a biosequence analysis using hidden Markov models. HMMER website (2010) <http://hmmer.wustl.edu/>.
56. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: A sequence logo generator. *Genome Research* 14: 1188–1190.
57. WebLogo: A sequence logo generator. WebLogo website (2010) <http://weblogo.berkeley.edu/>.
58. R Development Core Team (2009) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, R website (2010) <http://www.R-project.org>.
59. DeLano WL (2002) The PyMOL Molecular Graphics System. PyMOL website (2010) <http://www.pymol.org>.
60. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.